EAST Search History

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L1	15549	709/707 709/218 709/204 705/14	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 14:15
L2	105	craw$4 and (error near10 URL)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 14:15
L3	10	1 and L2	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 14:15
S1	63	web adj crawling	USPAT	OR	OFF	2005/04/14 09:37
S2	55	S1 and URL	USPAT	OR	OFF	2005/04/13 14:28
S3	14	"5999929"	USPAT	OR	OFF	2005/04/13 10:28
S4	48	"5974572"	USPAT	OR	OFF	2005/04/13 10:28
S5	47	("5974572").URPN.	USPAT	OR	OFF	2005/04/13 10:30
S6	9	"6253204"	USPAT	OR	OFF	2005/04/13 10:56
S7	1	"20020013782"	US-PGPUB; USPAT	OR	OFF	2005/04/13 10:56
S8	63	web adj crawling	USPAT	OR	OFF	2005/04/13 14:28
S9	55	S8 and URL	USPAT	OR	OFF	2005/04/13 14:28
S10	15	S9 and script	USPAT	OR	OFF	2005/04/13 14:28
S11	13	S10 and code	USPAT	OR	OFF	2005/04/13 14:28
S12	12	S11 and web adj page	USPAT	OR	OFF	2005/04/13 15:10
S13	0	S8 and script same URL same crawl$4	USPAT	OR	OFF	2005/04/13 15:10
S14	7	script same URL same crawl$4	US-PGPUB; USPAT; USOCR	OR	ON	2005/04/13 15:11
S15	0	("2004/0143787").URPN.	USPAT	OR	OFF	2005/04/13 15:38
S16	153	URL same crawl$4	USPAT	OR	OFF	2005/04/13 15:38
S17	147	S16 and @ad<"20020619"	USPAT	OR	OFF	2006/08/17 21:07
S18	35	S17 and script	USPAT	OR	OFF	2005/04/13 15:38
S19	1	"20020052928"	US-PGPUB; USPAT	OR	OFF	2005/04/14 09:37

EAST Search History

S20	20	script same crawl$4	US-PGPUB; USPAT	OR	OFF	2005/04/14 09:37
S21	16	S20 and @ad<"20020619"	US-PGPUB; USPAT	OR	OFF	2005/04/14 09:37
S22	40	identif$4 same script same code same web	USPAT	OR	OFF	2005/04/14 11:20
S23	39	S22 and @ad<"20020619"	USPAT	OR	OFF	2005/04/14 17:35
S24	0	"68571247"	USPAT	OR	OFF	2005/04/14 16:44
S25	1	"6857124"	USPAT	OR	OFF	2005/04/14 16:44
S26	40	identif$4 same script same code same web	USPAT	OR	OFF	2005/04/14 17:35
S27	39	S26 and @ad<"20020619"	USPAT	OR	OFF	2007/04/10 15:35
S28	12	S27 and notification	USPAT	OR	OFF	2005/04/14 18:10
S29	3106	search adj engine	USPAT	OR	OFF	2005/04/14 18:10
S30	5656	search$3 same engine	USPAT	OR	OFF	2005/04/14 18:10
S31	2608	S30 and web	USPAT	OR	OFF	2005/04/14 18:10
S32	21	web same page same load$4 same URL same script	USPAT	OR	OFF	2005/04/14 18:11
S33	6	S32 and (search adj engine)	USPAT	OR	OFF	2005/04/14 18:11
S34	6	S33 and @ad<"20020619"	USPAT	OR	OFF	2005/04/15 10:25
S35	4	"6424966"	USPAT	OR	OFF	2005/04/15 11:28
S36	0	"20020052928"	USPAT	OR	OFF	2005/04/15 11:28
S37	1	"20020052928"	US-PGPUB; USPAT	OR	OFF	2005/04/15 11:29
S38	1	"10064176"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 14:57
S39	1	"20020052928"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 15:12
S40	1	"20040143787" and execution	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 15:40
S41	1297	script$3 near10 URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 15:40

EAST Search History

S42	2070	execut$3 near10 URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 15:41
S43	179	S41 same S42	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 15:41
S44	141	S43 and (@ad<"20020619" @rlad<"20020619")	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 15:43
S45	10	S44 and 709/218.ccls.	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 15:46
S46	131	S44 not S45	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 15:46
S47	120	S46 and web	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50
S48	1297	script$3 near10 URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50
S49	2070	execut$3 near10 URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50
S50	179	S48 same S49	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50
S51	141	S50 and (@ad<"20020619" @rlad<"20020619")	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50

EAST Search History

S52	10	S51 and 709/218.ccls.	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50
S53	131	S51 not S52	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50
S54	120	S53 and web	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50
S55	120	S54	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:50
S56	3	S55 and crawler	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2005/12/28 17:51
S57	1	"20020052928"	USPAT	OR	OFF	2006/08/17 21:07
S58	0	"20020147637"	USPAT	OR	OFF	2006/08/17 21:07
S59	1	"20020147637"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/17 21:07
S60	14485	crawler$3	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/17 21:07
S61	12841635	@rlad<"20020619" @ad<"20020619"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/17 21:08
S62	10176	S61 and S60	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/17 21:08

EAST Search History

S63	716	S62 and URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/17 21:08
S64	678	URL same crawl$3	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/17 21:12
S65	454	S64 and S61	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/17 21:13
S66	124	S64 and S61 and spider$1	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/17 21:13
S67	2	"20020052928"	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S68	9713	709/224	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S69	29	707/14	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S70	87	web adj crawling	USPAT	OR	OFF	2006/08/18 17:03
S71	77	S70 and URL	USPAT	OR	OFF	2006/08/18 17:03
S72	21	"5999929"	USPAT	OR	OFF	2006/08/18 17:03
S73	67	"5974572"	USPAT	OR	OFF	2006/08/18 17:03
S74	66	("5974572").URPN.	USPAT	OR	OFF	2006/08/18 17:03
S75	10	"6253204"	USPAT	OR	OFF	2006/08/18 17:03
S76	1	"20020013782"	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S77	87	web adj crawling	USPAT	OR	OFF	2006/08/18 17:03
S78	77	S77 and URL	USPAT	OR	OFF	2006/08/18 17:03
S79	21	S78 and script	USPAT	OR	OFF	2006/08/18 17:03
S80	17	S79 and code	USPAT	OR	OFF	2006/08/18 17:03
S81	16	S80 and web adj page	USPAT	OR	OFF	2006/08/18 17:03
S82	0	S77 and script same URL same crawl$4	USPAT	OR	OFF	2006/08/18 17:03
S83	18	script same URL same crawl$4	US-PGPUB; USPAT; USOCR	OR	ON	2006/08/18 17:03

EAST Search History

S84	0	("2004/0143787").URPN.	USPAT	OR	OFF	2006/08/18 17:03
S85	211	URL same crawl$4	USPAT	OR	OFF	2006/08/18 17:03
S86	193	S85 and @ad<"20020619"	USPAT	OR	OFF	2006/08/18 17:03
S87	46	S86 and script	USPAT	OR	OFF	2006/08/18 17:03
S88	2	"20020052928"	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S89	39	script same crawl$4	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S90	20	S89 and @ad<"20020619"	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S91	62	identif$4 same script same code same web	USPAT	OR	OFF	2006/08/18 17:03
S92	58	S91 and @ad<"20020619"	USPAT	OR	OFF	2006/08/18 17:03
S93	0	"68571247"	USPAT	OR	OFF	2006/08/18 17:03
S94	1	"6857124"	USPAT	OR	OFF	2006/08/18 17:03
S95	62	identif$4 same script same code same web	USPAT	OR	OFF	2006/08/18 17:03
S96	58	S95 and @ad<"20020619"	USPAT	OR	OFF	2006/08/18 17:03
S97	16	S96 and notification	USPAT	OR	OFF	2006/08/18 17:03
S98	4235	search adj engine	USPAT	OR	OFF	2006/08/18 17:03
S99	7347	search$3 same engine	USPAT	OR	OFF	2006/08/18 17:03
S100	3642	S99 and web	USPAT	OR	OFF	2006/08/18 17:03
S101	29	web same page same load$4 same URL same script	USPAT	OR	OFF	2006/08/18 17:03
S102	8	S101 and (search adj engine)	USPAT	OR	OFF	2006/08/18 17:03
S103	6	S102 and @ad<"20020619"	USPAT	OR	OFF	2006/08/18 17:03
S104	7	"6424966"	USPAT	OR	OFF	2006/08/18 17:03
S105	1	"20020052928"	USPAT	OR	OFF	2006/08/18 17:03
S106	2	"20020052928"	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S107	1	"10064176"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03

EAST Search History

S108	2	"20020052928"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S109	1	"20040143787" and execution	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S110	1487	script$3 near10 URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S111	2379	execut$3 near10 URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S112	204	S110 same S111	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S113	157	S112 and (@ad<"20020619" @rlad<"20020619")	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S114	11	S113 and 709/218.ccls.	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S115	146	S113 not S114	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S116	135	S115 and web	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S117	1487	script$3 near10 URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03

EAST Search History

S118	2379	execut$3 near10 URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S119	204	S117 same S118	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S120	157	S119 and (@ad<"20020619" @rlad<"20020619")	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S121	11	S120 and 709/218.ccls.	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S122	146	S120 not S121	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S123	135	S122 and web	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S124	135	S123	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S125	3	S124 and crawler	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S126	1	"20020052928"	USPAT	OR	OFF	2006/08/18 17:03
S127	0	"20020147637"	USPAT	OR	OFF	2006/08/18 17:03
S128	1	"20020147637"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03

EAST Search History

S129	14485	crawler$3	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S130	12841635	@rlad<"20020619" @ad<"20020619"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S131	10176	S130 and S129	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S132	716	S131 and URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S133	678	URL same crawl$3	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S134	454	S133 and S130	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S135	124	S133 and S130 and spider$1	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2006/08/18 17:03
S136	2	"20020052928"	US-PGPUB; USPAT	OR	OFF	2006/08/18 17:03
S137	1	"20020019851"	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2007/04/10 15:34
S138	51	(script near2 code) with URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2007/04/10 15:37

EAST Search History

S139	5	S138 and (spider crawl$2)	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2007/04/10 15:37
S140	722910	@rlad<"20020619" and @ad<"20020619"	USPAT	OR	OFF	2007/04/10 15:35
S141	0	S140 and S139	USPAT	OR	OFF	2007/04/10 15:35
S142	857194	@rlad<"20020619" and @ad<"20020619"	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/04/10 15:36
S143	0	S142 and S139	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2007/04/10 15:36
S144	153	(script near2 code) same URL	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2007/04/10 15:37
S145	17	S144 and (spider crawl$2)	US-PGPUB; USPAT; USOCR; EPO; JPO; IBM_TDB	OR	ON	2007/04/10 15:37
S146	0	craw$4 and "Not Found"	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 10:44
S147	7684	craw$4 and error	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 10:46
S148	13060929	@ad<"20020619" @rlad<"20020619"	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 10:45

EAST Search History

S149	4891	S147 and S148	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 10:45
S150	105	craw$4 and (error near10 URL)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 10:46
S151	14	craw$4 and ((error near10 URL) same display$3)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 10:50
S152	12	S151 and S148	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; IBM_TDB	OR	ON	2007/11/30 14:15

Google

web crawler and URL Search Advanced Search Preferences

The "AND" operator is unnecessary -- we include all search terms by default. [details]

Web Results 1 - 10 of about 245,000 for web crawler and URL. (0.23 seconds)

WebCrawler Web Search
Offers a single source to search the **Web**, images, audio, video, news from Google, Yahoo!, MSN, Ask and many more search engines.
www.**webcrawler**.com/**WebCrawler**/Submit**URL**S.html - 23k - Cached - Similar pages

Url Definition - News - **WebCrawler**
News search results for **Url** Definition from **WebCrawler** Metasearch.
www.**webcrawler**.com/.../qkw=**Url**%20Definition/rfcp=RightNav/rfcid=302349/_ice**Url**Flag=11?_Ice**Url**=true - 82k - Cached - Similar pages
[More results from www.webcrawler.com]

About Ask.com: Webmasters
The **crawler** excludes some URLs if it has downloaded a sufficient number from the **Web** site or if it appears that the **URL** might be a duplicate of another **URL** ...
about.ask.com/en/docs/about/**web**masters.shtml - 22k - Cached - Similar pages

Url Definition - **Web** - **WebCrawler**
Web search results for **Url** Definition from **WebCrawler** Metasearch.
msxml.**webcrawler**.com/.../rfcid=407/qcat=**Web**/qkw=**Url**%20Definition/qcoll=Relevance/_ice**Url**Flag=11?_Ice**Url**=true - 80k - Cached - Similar pages

Definition of **Url** - **Web** - **WebCrawler**
Web search results for Definition of **Url** from **WebCrawler** Metasearch.
msxml.**webcrawler**.com/.../zoom=off/bepersistence=true/qi=21/qk=20/page=2/_ice**Url**Flag=11?_Ice**Url**=true - 84k - Cached - Similar pages
[More results from msxml.webcrawler.com]

Spider **Web** - **Web** spider, website **crawler**, **URL** extractor **Web** ...
Spider **Web** - **Web** spider, website **crawler**, **URL** extractor **Web** Extract with **Web** Spider / **Web Crawler**.
spider-**web**.qarchive.org/ - 24k - Cached - Similar pages

Tests of **URL** connections with the **Web crawler**
After you specify URLs for the **Web crawler** to crawl, you can test the configuration of the crawling rules.
publib.boulder.ibm.com/.../v8r3/topic/com.ibm.**web**sphere.ii.esearch.ad.doc/administering/iiysac**web**test.htm - 9k - Cached - Similar pages

FAST - Enterprise Search

At FAST we find needles in haystacks. In fact, digitally speaking, we find needles in trillions of far-flung haystacks. We help people come to confident ...

Sponsored Links

Domain Names
Register names at NetworkSolutions®
24/7 award winning customer support
NetworkSolutions.com

Custom **Web Crawler**
Crawl, Sort, Extract, Mine
Start Crawling in 3 Minutes
www.velocityscape.com

Web Spider Software
Extract **web** content **and** metadata from websites into your database
www.newprosoft.com

Web Crawler
Setup **web** crawling **and** extraction with GUI driven tool in minutes
www.sundewsoft.com

www.fastsearch.com/ - 30k - Cached - Similar pages

The **Web** Robots Pages

It has also been open for discussion on the Technical World Wide **Web** mailing ... This file must be accessible via HTTP on the local **URL** " /robots.txt ". ...
www.robotstxt.org/wc/robots.html - 12k - Cached - Similar pages

A **Web Crawler** in Perl

The **URL** Queue. Once the spider has downloaded the HTML source for a **web** page, ...
Re: A **Web Crawler** in Perl. On May 2nd, 2002 Anonymous says:. Hi Mike; ...
www.linuxjournal.com/article/2200 - 41k - Cached - Similar pages

1 2 3 4 5 6 7 8 9 10 **Next**

Download Google Pack: free essential software for your PC

---

web crawler and URL Search

Search within results | Language Tools | Search Tips | Dissatisfied? Help us improve

---

©2007 Google - Google Home - Advertising Programs - Business Solutions - About Google

PORTAL
USPTO
Subscribe (Full Service) Register (Limited Service, Free) Login
Search: The ACM Digital Library The Guide
web crawler and URL
SEARCH
THE ACM DIGITAL LIBRARY
Feedback Report a problem Satisfaction survey

Terms used: **web crawler** and **URL** Found **8,983** of **215,481**

Sort results by: relevance
Display results: expanded form
Save results to a Binder
Search Tips
Open results in a new window
Try an Advanced Search
Try this search in The ACM Guide

Results 1 - 20 of 200 Result page: **1** 2 3 4 5 6 7 8 9 10 next
Best 200 shown Relevance scale

1 Web crawling and measurement: Efficient URL caching for world wide web crawling
Andrei Z. Broder, Marc Najork, Janet L. Wiener
May 2003 **Proceedings of the 12th international conference on World Wide Web WWW '03**
Publisher: ACM Press
Full text available: pdf(174.37 KB) Additional Information: full citation, abstract, references, citings, index terms

Crawling the web is deceptively simple: the basic algorithm is (a) Fetch a page (b) Parse it to extract all linked URLs (c) For all the URLs not seen before, repeat (a)-(c). However, the size of the web (estimated at over 4 billion pages) and its rate of change (estimated at 7% per week) move this plan from a trivial programming exercise to a serious algorithmic and system design challenge. Indeed, these two factors alone imply that for a reasonably fresh and complete crawl of the web, step (a) ...

Keywords: URL caching, caching, crawling, distributed crawlers, web crawlers, web graph models

2 Web engineering: Evaluation of crawling policies for a web-repository crawler
Frank McCown, Michael L. Nelson
August 2006 **Proceedings of the seventeenth conference on Hypertext and hypermedia HYPERTEXT '06**
Publisher: ACM Press
Full text available: pdf(482.40 KB) Additional Information: full citation, abstract, references, citings, index terms

We have developed a web-repository crawler that is used for reconstructing websites when backups are unavailable. Our crawler retrieves web resources from the Internet Archive, Google, Yahoo and MSN. We examine the challenges of crawling web repositories, and we discuss strategies for overcoming some of these obstacles. We propose three crawling policies which can be used to reconstruct websites. We evaluate the effectiveness of the policies by reconstructing 24 websites and comparing the result ...

Keywords: crawler policy, digital preservation, search engine, website reconstruction

3 LSCrawler: A Framework for an Enhanced Focused Web Crawler Based on Link

Semantics

M. Yuvarani, N. Ch. S. N. Iyengar, A. Kannan

December 2006 **Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence WI '06**

Publisher: IEEE Computer Society

Full text available: pdf(191.22 KB) Additional Information: full citation, abstract, index terms

The traditional process of focused web crawler is to harvest a collection of web documents that are focused on the topical subspaces. The intricacy of focused crawlers is identifying the next most important and relevant link to follow. Focused Crawlers mostly rely on probabilistic models for predicting the relevancy of the documents. The Web documents are well characterized by the hypertext and the hypertext can be used to determine the relevance of the document to the search domain. The semanti ...

4 On the design of a learning crawler for topical resource discovery

Charu C. Aggarwal, Fatima Al-Garawi, Philip S. Yu

July 2001 **ACM Transactions on Information Systems (TOIS)**, Volume 19 Issue 3

Publisher: ACM Press

Full text available: pdf(324.39 KB) Additional Information: full citation, abstract, references, citings, index terms

In recent years, the World Wide Web has shown enormous growth in size. Vast repositories of information are available on practically every possible topic. In such cases, it is valuable to perform topical resource discovery effectively. Consequently, several new ideas have been proposed in recent years; among them a key technique is focused crawling which is able to crawl particular topical portions of the World Wide Web quickly, without having to explore all web pages. In this paper, we propose ...

Keywords: Crawling, World Wide Web

5 Evaluating topic-driven web crawlers

Filippo Menczer, Gautam Pant, Padmini Srinivasan, Miguel E. Ruiz

September 2001 **Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '01**

Publisher: ACM Press

Full text available: pdf(210.09 KB) Additional Information: full citation, abstract, references, citings, index terms

Due to limited bandwidth, storage, and computational resources, and to the dynamic nature of the Web, search engines cannot index every Web page, and even the covered portion of the Web cannot be monitored continuously for changes. Therefore it is essential to develop effective crawling strategies to prioritize the pages to be indexed. The issue is even more important for topic-specific search engines, where crawlers must make additional decisions based on the relevance of visited pages. ...

Keywords: InfoSpiders, PageRank, Web information retrieval, best-first search, focused crawlers, performance metrics, topic driven crawling

6 Intelligent crawling on the World Wide Web with arbitrary predicates

Charu C. Aggarwal, Fatima Al-Garawi, Philip S. Yu

April 2001 **Proceedings of the 10th international conference on World Wide Web WWW '01**

Publisher: ACM Press

Full text available: pdf(272.60 KB) Additional Information: full citation, references, citings, index terms

Keywords: World Wide Web, crawling, querying

7 Web resource crawling and searching: Lazy preservation: reconstructing websites by crawling the crawlers

Frank McCown, Joan A. Smith, Michael L. Nelson

November 2006 **Proceedings of the 8th annual ACM international workshop on Web information and data management WIDM '06**

Publisher: ACM

Full text available: pdf(720.52 KB) Additional Information: full citation, abstract, references, index terms

Backup of websites is often not considered until after a catastrophic event has occurred to either the website or its webmaster. We introduce "lazy preservation" -- digital preservation performed as a result of the normal operation of web crawlers and caches. Lazy preservation is especially suitable for third parties; for example, a teacher reconstructing a missing website used in previous classes. We evaluate the effectiveness of lazy preservation by reconstructing 24 websites of varying sizes ...

Keywords: cached resources, digital preservation, recovery, search engine

8 Tracking the changes of dynamic web pages in the existence of URL rewriting

Ping-Jer Yeh, Jie-Tsung Li, Shyan-Ming Yuan

November 2006 **Proceedings of the fifth Australasian conference on Data mining and analystics - Volume 61 AusDM '06**

Publisher: Australian Computer Society, Inc.

Full text available: pdf(402.54 KB) Additional Information: full citation, abstract, references

Crawlers in a knowledge management system need to collect and archive documents from websites, and also track the change status of these documents. However, the existence of URL rewriting mechanism raises a page tracking problem since the URLs of a pair of dynamic page instances obtained during different sessions will no longer be the same. This paper proposes a series of algorithms in a bottom-up manner to find the corresponding pairs of dynamic page instances, and then to judge the change s ...

Keywords: HTTP session, URL rewriting, crawler, string matching

9 Crawler-Friendly Web Servers

Onn Brandman, Junghoo Cho, Hector Garcia-Molina, Narayanan Shivakumar

September 2000 **ACM SIGMETRICS Performance Evaluation Review**, Volume 28 Issue 2

Publisher: ACM Press

Full text available: pdf(513.04 KB) Additional Information: full citation, abstract, citings, index terms

In this paper we study how to make web servers (e.g., Apache) more crawler friendly. Current web servers offer the same interface to crawlers and regular web surfers, even though crawlers and surfers have very different performance requirements. We evaluate simple and easy-to-incorporate modifications to web servers so that there are significant bandwidth savings. Specifically, we propose that web servers export meta-data archives decribing their content.

10 Crawling: Parallel crawlers

Junghoo Cho, Hector Garcia-Molina

May 2002 **Proceedings of the 11th international conference on World Wide Web**

WWW '02
Publisher: ACM Press

Full text available: pdf(230.70 KB) Additional Information: full citation, abstract, references, citings, index terms

In this paper we study how we can design an effective parallel crawler. As the size of the Web grows, it becomes imperative to parallelize a crawling process, in order to finish downloading pages in a reasonable amount of time. We first propose multiple architectures for a parallel crawler and identify fundamental issues related to parallel crawling. Based on this understanding, we then propose metrics to evaluate a parallel crawler, and compare the proposed architectures using 40 million pages ...

Keywords: parallelization, web crawler, web spider

11 Web retrieval II (IR): Designing clustering-based web crawling policies for search engine crawlers

Qingzhao Tan, Prasenjit Mitra, C. Lee Giles
November 2007 **Proceedings of the sixteenth ACM conference on Conference on information and knowledge management CIKM '07**
Publisher: ACM

Full text available: pdf(270.20 KB) Additional Information: full citation, abstract, references, index terms

The World Wide Web is growing and changing at an astonishing rate. Web information systems such as search engines have to keep up with the growth and change of the Web. Due to resource constraints, search engines usually have difficulties keeping the local database completely synchronized with the Web. In this paper, we study how tomake good use of the limited system resource and detect as many changes as possible. Towards this goal, a crawler for the Web search engine should be able to predi ...

Keywords: clustering, incremental crawler, refresh policy, sampling, web search engine

12 Web 2: Structure-driven crawler generation by example

Márcio L. A. Vidal, Altigran S. da Silva, Edleno S. de Moura, João M. B. Cavalcanti
August 2006 **Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '06**
Publisher: ACM Press

Full text available: pdf(639.17 KB) Additional Information: full citation, abstract, references, index terms

Many Web IR and Digital Library applications require a crawling process to collect pages with the ultimate goal of taking advantage of useful information available on Web sites. For some of these applications the criteria to determine when a page is to be present in a collection are related to the page content. However, there are situations in which the inner structure of the pages provides a better criteria to guide the crawling process than their content. In this paper, we present a structure- ...

Keywords: digital libraries, tree edit distance, web crawlers

13 Characteristics of streaming media stored on the Web

Mingzhe Li, Mark Claypool, Robert Kinicki, James Nichols
November 2005 **ACM Transactions on Internet Technology (TOIT)**, Volume 5 Issue 4
Publisher: ACM Press

Full text available: pdf(936.68 KB) Additional Information: full citation, abstract, references, citings, index terms

Despite the growth in multimedia, there have been few studies that focus on

characterizing streaming audio and video stored on the Web. This investigation used a customized Web crawler to traverse 17 million Web pages from diverse geographic locations and identify nearly 30,000 streaming audio and video clips available for analysis. Using custom-built extraction tools, these streaming media objects were analyzed to determine attributes such as media type, encoding format, playout duration, bitra ...

Keywords: Apple QuickTime, Microsoft Windows Media Player, RealNetworks RealPlayer, long-tailed, multimedia, self-similarity, streaming

14 Web search 1: Topic-oriented collaborative crawling

Chiasen Chung, Charles L. A. Clarke

November 2002 **Proceedings of the eleventh international conference on Information and knowledge management CIKM '02**

Publisher: ACM Press

Full text available: pdf(179.28 KB) Additional Information: full citation, abstract, references, citings, index terms

A major concern in the implementation of a distributed Web crawler is the choice of a strategy for partitioning the Web among the nodes in the system. Our goal in selecting this strategy is to minimize the overlap between the activities of individual nodes. We propose a topic-oriented approach, in which the Web is partitioned into general subject areas with a crawler assigned to each. We examine design alternatives for a topic-oriented distributed crawler, including the creation of a Web page cl ...

Keywords: distributed systems, text categorization, web crawling

15 I/O-conscious data preparation for large-scale web search engines

Maxim Lifantsev, Tzi-cker Chiueh

August 2002 **Proceedings of the 28th international conference on Very Large Data Bases - Volume 28 VLDB '2002**

Publisher: VLDB Endowment

Full text available: pdf(292.60 KB) Additional Information: full citation, abstract, references, index terms

Given that commercial search engines cover billions of web pages, efficiently managing the corresponding volumes of disk-resident data needed to answer user queries quickly is a formidable data manipulation challenge. We present a general technique for efficiently carrying out large sets of simple transformation or querying operations over external-memory data tables. It greatly reduces the number of performed disk accesses and seeks by maximizing the temporal locality of data access and orga ...

16 Models/measurements of traffic/web systems: Web graph analyzer tool

Konstantin Avrachenkov, Danil Nemirovsky, Natalia Osipova

October 2006 **Proceedings of the 1st international conference on Performance evaluation methodolgies and tools valuetools '06**

Publisher: ACM Press

Full text available: pdf(160.03 KB) Additional Information: full citation, abstract, references, index terms

We present the software tool "Web Graph Analyzer". This tool is designed to perform a comprehensive analysis of the Web Graph structure. By Web Graph we mean a graph whose vertices are Web pages and whose edges are hyper-links. With the help of the Web Graph Analyzer we can study the local graph characteristics such as numbers and sets of incoming/outgoing links to/from a given page, the page level relative to a given root page, and the global graph characteristics such as PageRank, Giant Strong ...

Keywords: PageRank, World Wide Web (WWW), connectivity, crawler, graph theory,

software tool, web graph

17 Posters: Distributed location aware web crawling

Odysseas Papapetrou, George Samaras

May 2004 **Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters WWW Alt. '04**

Publisher: ACM Press

Full text available: pdf(29.33 KB) Additional Information: full citation, abstract, references, index terms

Distributed crawling has shown that it can overcome important limitations of the today's crawling paradigm. However, the optimal benefits of this approach are usually limited to the sites hosting the crawler. In this work, we propose a location-aware method, called IPMicra, that utilizes an IP address hierarchy, and allows crawling of links in a near optimal location aware manner.

Keywords: distributed web crawling, location aware web crawling

18 Crawling the web: Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method

Jialun Qin, Yilu Zhou, Michael Chau

June 2004 **Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries JCDL '04**

Publisher: ACM Press

Full text available: pdf(214.74 KB) Additional Information: full citation, abstract, references, citings, index terms

Collecting domain-specific documents from the Web using focused crawlers has been considered one of the most important strategies to build digital libraries that serve the scientific community. However, because most focused crawlers use local search algorithms to traverse the Web space, they could be easily trapped within a limited sub-graph of the Web that surrounds the starting URLs and build domain-specific collections that are not comprehensive and diverse enough to scientists and researcher ...

Keywords: digital libraries, domain-specific collection building, focused crawling, meta-search, web search algorithm

19 Web ranking and classification: Efficient, automatic web resource harvesting

Michael L. Nelson, Joan A. Smith, Ignacio Garcia del Campo

November 2006 **Proceedings of the 8th annual ACM international workshop on Web information and data management WIDM '06**

Publisher: ACM

Full text available: pdf(703.63 KB) Additional Information: full citation, abstract, references, index terms

There are two problems associated with conventional web crawling techniques: a crawler cannot know if all resources at a non-trivial web site have been discovered and crawled ("the counting problem") and the human-readable format of the resources are not always suitable for machine processing ("the representation problem"). We introduce an approach that solves these two problems by implementing support for both the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and MPEG-21 D ...

Keywords: OAI-PMH, mod_oai, web crawling

20 Stanford WebBase components and applications

Junghoo Cho, Hector Garcia-Molina, Taher Haveliwala, Wang Lam, Andreas Paepcke, Sriram Raghavan, Gary Wesley
May 2006 **ACM Transactions on Internet Technology (TOIT)**, Volume 6 Issue 2
Publisher: ACM Press
Full text available: pdf(609.18 KB) Additional Information: full citation, abstract, references, index terms

We describe the design and performance of WebBase, a tool for Web research. The system includes a highly customizable crawler, a repository for collected Web pages, an indexer for both text and link-related page features, and a high-speed content distribution facility. The distribution module enables researchers world-wide to retrieve pages from WebBase, and stream them across the Internet at high speed. The advantage for the researchers is that they need not all crawl the Web before beginning t ...

Keywords: WebBase Web crawler, distribution, hyperlink indexing, site crawling

Results 1 - 20 of 200 Result page: **1** 2 3 4 5 6 7 8 9 10 next

The ACM Portal is published by the Association for Computing Machinery. Copyright © 2007 ACM, Inc.
Terms of Usage Privacy Policy Code of Ethics Contact Us

Useful downloads: Adobe Acrobat QuickTime Windows Media Player Real Player

Results for "((web crawler and url)<in>metadata)"
Your search matched 7 of 1692897 documents.
A maximum of 100 results are displayed, 25 to a page, sorted by Relevance in Descending order.

New [Beta] Application Notes POWERED BY GLOBALSPEC

Modify Search

((web crawler and url)<in>metadata) Search

Check to search only within this results set

Display Format: Citation Citation & Abstract

» Search Options

View Session History

New Search

» Key

IEEE JNL	IEEE Journal or Magazine
IET JNL	IET Journal or Magazine
IEEE CNF	IEEE Conference Proceeding
IET CNF	IET Conference Proceeding
IEEE STD	IEEE Standard

IEEE/IET Books Educational Courses A

IEEE/IET journals, transactions, letters, magazines, conference proceedings, and

view selected items Select All Deselect All

1. **Ontology-based Web crawler**
Ganesh, S.; Jayaraj, M.; Kalyan, V.; SrinivasaMurthy; Aghila, G.;
Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2(Conference on
Volume 2, 2004 Page(s):337 - 341 Vol.2
Digital Object Identifier 10.1109/ITCC.2004.1286658
AbstractPlus | Full Text: PDF(1400 KB) IEEE CNF
Rights and Permissions

2. **CINDI Robot: an Intelligent Web Crawler Based on Multi-level Inspection**
Chen, Rui; Desai, Bipin C.; Zhou, Cong;
Database Engineering and Applications Symposium, 2007. IDEAS 2007. 11th
6-8 Sept. 2007 Page(s):93 - 101
Digital Object Identifier 10.1109/IDEAS.2007.4318093
AbstractPlus | Full Text: PDF(265 KB) IEEE CNF
Rights and Permissions

3. **WebGuard: Web based adult content detection and filtering system**
Hammami, M.; Chahir, Y.; Chen, L.;
Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Confe
13-17 Oct. 2003 Page(s):574 - 578
AbstractPlus | Full Text: PDF(374 KB) IEEE CNF
Rights and Permissions

4. **The Design and Implementation of the Crawler-Inar**
Yu-Xin Ding; Xiao-Long Wang; Le-Bin Lin; Qi Zhang; Yong-Hui Wu;
Machine Learning and Cybernetics, 2006 International Conference on
Aug. 2006 Page(s):4527 - 4530
Digital Object Identifier 10.1109/ICMLC.2006.259171
AbstractPlus | Full Text: PDF(142 KB) IEEE CNF
Rights and Permissions

5. **AuToCrawler: an integrated system for automatic topical crawler**
Jyh-Jong Tsay; Chen-Yang Shih; Bo-Liang Wu;

Computer and Information Science, 2005. Fourth Annual ACIS International C
2005 Page(s):462 - 467
Digital Object Identifier 10.1109/ICIS.2005.33

AbstractPlus | Full Text: PDF(160 KB) IEEE CNF
Rights and Permissions

6. **A World Wide Web region-based image search engine**
Kompatsiaris, I.; Triantafyllou, E.; Strintzis, M.G.;
Image Analysis and Processing, 2001. Proceedings. 11th International Confe
26-28 Sept. 2001 Page(s):392 - 397
Digital Object Identifier 10.1109/ICIAP.2001.957041

AbstractPlus | Full Text: PDF(608 KB) IEEE CNF
Rights and Permissions

7. **A Kernel-based Algorithm for Multilevel Drawing Web Graphs**
Huang, Xiaodi; Lai, Wei; Zhang, Di; Huang, Mao Lin; Nguyen, Quang Vinh;
Computer Graphics, Imaging and Visualisation, 2007. CGIV '07
14-17 Aug. 2007 Page(s):454 - 459
Digital Object Identifier 10.1109/CGIV.2007.7

AbstractPlus | Full Text: PDF(398 KB) IEEE CNF
Rights and Permissions

Help Contact Us

Indexed by Inspec®

© Copyright 20